13

Two-way Tables of Counts

		Page
Introduction		13-1
13A.	Fitting and residuals	13-2
13B.	What if the factors have a before-after relation?	13-10
13C.	One factor ordered; unsymmetric treatment	13-17
13D.	Symmetric factors, one ordered	13-25
13E.	Both factors ordered; unsymmetric treatment	13-26
13F.	Both factors ordered; symmetric treatment desired	13-32
13G.	What to do with big cells -- and with prominent diagonals	13-36
13H.	How far have we come?	13-41

*****

13K.	Easy f-roots and f-logs with the slide rule	13-44
13L.	Sequenced factors may not be ordered	13-46
13M.	The treatment of mobility data	13-54
13N.	Symmetrical analyses with both factors ordered	13-61
13O.	Why does the kind of factor matter?	13-62
13P.	IEDA processor features	13-66

*****

13X.	Example - further analysis of the mobility data	13-67
13Y.	Problems	13-73

Location of tables and plots for 13

Number	Page
1	13-4
2	13-8
3	13-11
4	13-13
5	13-14
6	13-16
7	13-18
8	13-20
9	13-22
10	13-24
11	13-27
12	13-29
13	13-31
14	13-33
15	13-35
16	13-39
17	13-47
18	13-50
19	13-53
20	13-55
21	13-59
22	13-60
23	13-68
24	13-72
25	13-74
26	13-75
27	13-76

13

Two-way Tables of Counts

Counts arrayed in two-way tables are a relatively familiar sight. Tables of this sort serve many purposes. Any of them is likely to be called "a contingency table". While the central issue in this chapter is simple -- looking at two-way tables of counts to see if either factor seems to affect the distribution of cases over the levels or versions of the other -- we will spend much of our time in explaining why one example or another really involves a different purpose and ought to be approached in a specific way, often one of the ways we have already learned. In almost every case, however, whichever approach is to be chosen, the implicit standard to which the table of counts is to be referred is the same.

What then is this null situation, the implicit standard of reference for most two-way tables whose entries are of counts -- counts rather than any other kind of number? Consider two rare happenings that can -- but need not -- occur together. If one, A, happens 1% of the time, the other, B, 2%, and we have examined exactly 10,000 cases, the following 2x2 table is generally agreed to represent a complete lack of relationship of each event to the other.

	not A	A	(total)
not B:	9702	98	(9800)
B:	198	2	(200)
(total)	(9900)	(100)	((10,000))

Knowing about A tells us nothing about B -- whether we have A or not A, exactly two cases in 100 are B's. B tells us -- equally -- nothing about A. The characteristic of this null table is its multiplicative character -- multiplicative not additive. The fraction of cases in the 'A' - 'not B' cell, .0098 in the example, is equal to the result of multiplying the fraction in the 'A' column, .01 in the example, by the fraction in the 'not B' row, .98 in the example.

More generally, complete lack of relationship is taken to be characterized by

fraction of cases in cell

= (fraction in row)x(fraction in column)

It is this relation that almost always provides an implicit standard of comparison for a two-way table of counts. (Should the table be incomplete, the row and column fractions would allow for the missing counts.)

The crucial ways in which two-way tables of counts differ from the two-way tables of other responses that we have been analyzing are thus:

- as counts, we expect to want to take square roots.
- the implicit standard of reference is multiplicative, not additive.

The new elements in this chapter are associated with meeting these challenges, either singly or together.

13A. Fitting and residuals

Let us start with a classical example, used by Yule in his An Introduction to the Theory of Statistics (first edition 1910, tenth edition, revised 1936) and credited to

Ammon, Zur Anthropologie der Badener, concerning the hair and eye color of 6800 males in Baden. The data is given in the first panel of table 1.

Panel (b) gives the fitted counts based on the marginal totals. As shown by the example at the foot of the table, the working arithmetic is very simple. The reason behind it is almost equally simple, namely

$$\text{fraction with blue eyes} = \frac{2811}{6800}$$

$$\text{fraction with fair hair} = \frac{2829}{6800}$$

$$\text{product of these} = \frac{(2811)(2829)}{(6800)(6800)}$$

The last of these is the implicit reference fraction for the extreme NW cell. To change this from a fraction to a count, to make an implied reference count, we need to multiply this fraction by 6800. This cancels one factor of 6800 in the denominator and gives the form shown in the table's example.

In finding the square root corresponding to the observed we have used $\sqrt{4(\text{obs}) + 2}$ rather than $\sqrt{\text{obs}} + \sqrt{1+\text{obs}}$ (as in table 2 of chapter 11). This seems to be a greater divergence than it is. Consider the following figures:

count:	0	1	2	3	4
$\sqrt{\text{count}} + \sqrt{1+\text{count}}$:	1.000	2.414	3.146	3.732	4.236
$\sqrt{2 + 4(\text{count})}$:	1.414	2.449	3.162	3.742	4.243

Except for count = 1, no figure differs as much as in the first decimal place. (For count ≥ 4, no difference is more than 0.007.) Accordingly, we may as well say that we are

table 1 of chapter 13

Ammon's Data on Hair and Eye-color from 6800 Males in Baden

Eyes:	Hair Color: Fair	Brown	Black	Red	(T)
a) OBSERVED COUNTS					
Blue	1768	807	189	47	(2811)
Gray or Green	946	1387	746	53	(3132)
Brown	115	438	288	16	(857)
(T)	(2829)	(2632)	(1223)	(116)	((6800))
b) FITTED COUNTS: from marginal totals, see example					
Blue	1169	1088	506	48.0	(2811)
Gray or Green	1303	1212	563	53.4	(3131.4)
Brown	357	332	154	14.6	(857.6)
(T)	(2829)	(2632)	(1223)	(116.0)	((6800))
c) SQUARE ROOTS FOR OBSERVED: $(4(\text{obs}) + 2)^{\frac{1}{2}}$					
Blue	84.1	56.7	18.9	13.8	
Gray or Green	61.5	74.5	54.5	14.6	
Brown	21.5	42.0	34.0	8.1	
d) SQUARE ROOTS FOR FITTED: $(4(\text{fit}) + 1)^{\frac{1}{2}}$					
Blue	68.4	66.0	45.0	13.9	
Gray or Green	72.4	69.5	47.6	14.6	
Brown	37.8	36.4	24.8	7.7	
e) RESIDUALS: difference of last two					
Blue	15.7	-9.3	-26.1	-0.1	
Gray or Green	-10.9	5.0	6.9	0.0	
Brown	-16.3	5.6	+9.2	0.4	

Source: Ammon, Zur Anthropologie der Badener.

Examples: $1169 = \frac{(2811)(2829)}{(6800)}$

$$84.1 = \sqrt{4(1768) + 2} = (4(1768) + 2)^{\frac{1}{2}}$$

$$68.4 = \sqrt{4(1169) + 1} = (4(1169) + 1)^{\frac{1}{2}}$$

$$15.7 = 84.1 - 68.4$$

using $\sqrt{2 + 4(\text{count})}$, at least except when count = 0, when we use 1. (A reason for preferring this form in the present situation will shortly become clear.)

When we square-root the fitted values, we do not use quite the same formula. Instead of using $\sqrt{2 + 4(\text{count})}$ we use $\sqrt{1 + 4(\text{count})}$. There is a reason for this seemingly arbitrary action. Since an adequate explanation would take us far outside the scope of the present account, we shall not try to give one.

((For those who must have a hint, it will have to be enough to know that the reason for such a distinction between how we square-root an observed value and how we square-root a fitted value is closely linked to the downward curvature of the square root. If we turn to plots 2 and 4 of chapter 2, for instance, and consider where the curve for the (matched) square root will have to fall, we see that it must fall below the straight line $y = x$ at both ends. Thus it will be generally somewhat lower. It is consequences of this that we are compensating for when we use "+1" for fit as against "+2" for observed.))

So much for the arithmetic and the reasons for its use. What about the results of the analysis? In particular, what about the residuals? We see numbers of varying size among them. Those for red hair are small, both relatively to the others and relatively to our previous experience. We have seen, in other examples where this sort of square root was used, many residuals of size between 0 and 1, fewer of sizes between 1 and 2 -- instances of values larger than 2 that

proved -- or seemed likely -- to be accidental were quite rare. Values like 5, 6 and 9 are thus quite substantial, while values like 15, 16, and 25 are very large indeed. Since residuals of these sizes appear, we are almost forced to believe that, in Baden, eye color and hair color ARE associated.

Especially in tables with more rows, or more columns, or both, the fitting procedure underlying the analysis we have just done has both advantages and disadvantages. In particular:

- if each cell contains a rather well-behaved count, then the marginal totals are both natural and effective to use.
- if one (or a few) cells is unusual enough (usually by being unduly large) it can dominate the corresponding marginal totals, and thus distort the fit.

There is a real place for a fitting procedure based on marginal totals; there is another, equally real place for a fitting procedure that is more resistant to one -- or a few -- erratic cells. When we start to cast around for such a procedure, we can recall that we have already set up a resistant procedure for analyzing a two-way table of responses -- one that depends on comparison with additive effects -- namely the use of midmean differences. It is natural to want to try to use the same approach here.

We have a multiplicative implied reference -- how can we make it an additive one? As always, the answer is "Take logs". If the cell counts fit a multiplicative model perfectly, their logs will fit an additive one! So let us try using them. (So long as there are no zero values, plain logs will do.) If there are zeroes, we can

either use log(count + 0.1) or log(count + 0.25), or, perhaps, if there are not too many zero entries, just leave out any entry for such a cell. (This last is not really satisfactory, however.)

The tolerance of analyses by midmean differences is great, which helps us greatly. Not only can we live with a small fraction of missing values, or with a somewhat larger fraction of (more or less arbitrarily) modified logs, but midmean differences do much to reduce trouble from the fact that cells with small counts tend to have less firmly fixed logarithms than do cells with large counts.

Once we combine their use with the use of midmean differences, logs look reasonably attractive. Table 2 sets out an analysis using them. In order to show the effect of including relatively small cells, table 2 first analyzes the data omitting the column for red hair, and then goes through all the calculations again including this column. The difference between the two analyses amounts to 0.8 in three of the nine residuals in common, and to this same amount in one stripe adjustment (once we allow for a constant shift in stripe adjustments). Clearly the small cells made little difference here.

A verbal summary of the information in the residuals runs roughly as follows:

- Fair hair and blue eyes go together about three times as often as indicated by the stripe adjustments.
- Black hair and blue eyes go together about 1/2 as often as would be computed from the stripe adjustments.

table 2 of chapter 13

Ammon on the Anthropology of the Badeners,
This Time Analyzed by Logs

Eyes	Fair hair	Brown hair	Black hair	Red hair
a) RAW COUNTS (again, see table 1):				
Blue	1768	867	189	47
Grey or Green	946	1387	746	53
Brown	115	438	288	16
b′) LOG COUNTS ($\log_{10}$(count), omitting red hair):				
Blue	3.25	2.94	2.28	
Grey or Green	2.98	3.14	2.87	
Brown	2.06	2.64	2.46	
c′) ELEMENTARY ANALYSIS (using midmean* differences):				
Blue	.47	.00	-.39	.00
Grey or Green	.00	.00	-.00	.20
Brown	-.42	.00	.09	-.30
	.00	+.16	-.11	2.78
d′) RESIDUALS AS % (undoing logarithms, rounding):				
Blue	300%	100%	40%	
Grey or Green	100%	100%	100%	
Brown	40%	100%	125%	

e′) STRIPE ADJUSTMENTS AS (multiple) RATIOS (undoing logarithms):

Blue/G-or-G/Brown = 1.00/1.58/0.50

Fair/Brown/Black = 1.00/1.45/0.78

b″) LOG COUNTS ($\log_{10}$count, including red hair):				
Blue	3.25	2.94	2.28	1.67
Green or Grey	2.98	3.14	2.87	1.72
Brown	2.06	2.64	2.46	1.20

(Cont'd)

table 2 (Cont'd)

c") ELEMENTARY ANALYSIS (using midmean* differences):

Eyes	Fair hair	Brown hair	Black hair	Red hair	
Blue	.39	-.08	-.47	.07	.00
Green or Grey	.00	.00	.00	.00	.12
Brown	-.41	.01	.10	-.01	-.39
	.06	.22	-.05	-1.20	2.80

d") RESIDUALS AS % (undoing logarithms, rounding):

Blue	250%	85%	35%	120%
Green or Grey	100%	100%	100%	100%
Brown	40%	100%	130%	10%

e") STRIPE ADJUSTMENTS AS (multiple) RATIOS (undoing logarithms):

Blue/G-or-G/Brown = 1.00/1.32/0.41

Fair/Brown/Black/Red = 1.15/1.66/0.89/0.06

*Midmeans of 4, calculated as usual; midmeans of 3, taken as medians.

Examples: $3.25 = \log_{10} 1768.$

$3.25 = .47 + .00 + .00 + 2.78.$

$.47 = \log_{10} 3.0,\ 3.0 = 300\%.$

$.16 = \log_{10} 1.58.$

$3.25 = \log_{10} 1768.$

$3.25 = .39 + .06 + .00 + 2.80.$

$.39 = \log_{10} 2.5,\ 2.5 = 250\%.$

$.06 = \log_{10} 1.15.$

- Fair hair and brown eyes go together about 1/3 as often as would be computed from the stripe adjustments.

If we do not need to worry about the fluctuations and irregularities that belong to our specific body of data, the logarithmic analysis does better in getting at the root of the matter. If fluctuations and irregularities had been so large -- as they often are -- that we would be lucky to have modest confidence in pointing out some cells as "high" and others as "low", the square-root analysis would have done better in getting to those roots of the matter that might be reachable.

This contrast between the usefulnesses of square-root and logarithmic methods, when both are applicable, is by no means confined to contingency tables. Although we did not point this out in Chapter 11 (as we should) the same is true in dealing with counts. Square roots help us in dealing with irregularities and fluctuations. Logs help us once the data is massive enough to be decisive. (As rat mites show, logs sometimes help us in dealing with irregularities and fluctuations.)

13B. <u>What if the factors have a before-after relation?</u>

Yule condensed data of Sir Francis Galton (as reported by Karl Pearson 1900) on the eye-color of fathers and sons to the form given in table 3. The logarithmic analysis given in panel (c) of that table shows that the four cells along the diagonal where father and son have the same eye color contain two or three times as many cases as the stripe

table 3 of chapter 13

Frequencies of Different Combinations of Eye Color in Father and Son

1 = Blue; 2 = Blue green or grey;

3 = Dark grey or hazel; 4 = Brown

Son's eyes	Father's eyes 1	2	3	4
a) DATA				
1	194	70	41	30
2	83	124	41	36
3	25	34	55	23
4	56	36	43	109
b) LOGARITHMS: ($\log_{10}$)				
1	2.29	1.85	1.61	1.48
2	1.92	2.09	1.61	1.56
3	1.40	1.53	1.74	1.36
4	1.75	1.56	1.63	2.04
c) ELEMENTARY ANALYSIS: (of logs, by midmean differences)					
1	.45	.04	-.07	-.11	.04
2	.04	.24	-.11	-.07	.08
3	-.17	-.01	.33	+.04	-.23
4	.00	-.16	.04	.54	-.05
	.10	.07	-.06	-.15	1.70

Examples: $2.29 = \log_{10} 194$.

$2.29 = .45 + .04 + .10 + 1.70$.

Source: Karl Pearson 1900. Phil. Trans. Roy. Soc. London, Ser. A 195: 138.

adjustment account for. Clearly this analysis has been effective in doing what it was asked to do. We are free, however, to ask whether what it was asked to do was the best choice.

The logarithmic analysis treats both factors -- father's eye color and son's eye color in exactly the same way -- it is truly symmetrical. For tables 1 and 2, which interrelate hair color with eye color in a single person, there seems to be no natural objection to such symmetry. In the present example this is not the case -- there are arguments against symmetry.

One strong argument can begin by pointing out that you cannot have sons without fathers, and go on to argue that we ought therefore to look at four distributions of son's eye color, one for each parent's eye color. If we accept this argument, the natural presentation involves:

- a set of rootobar plots.
- suspension of the rootobars from reference values combining the stripe-adjustment distribution of son's eye color and the observed number of fathers of a given eye color.

Table 4 shows the computations; plot 5 pictures the results. Again we see that agreement between father and son is more common than the implicit reference would suggest.

In this particular example, the argument that we should consider the data as four distributions -- one for each father's eye color -- was strong but not absolutely clinching. There is a counterargument, focussed upon the fact

table 4 of chapter 13

Calculations of Suspensions for Rootobars Expressing the Counts of Table 3(a)

a) BASE DISTRIBUTION:

Son's eyes	Row adjustments (from table 3(c)) as logs	as ratios	as fractions
1	.04	1.10	29.0%
2	.08	1.20	31.8%
3	-.23	.59	15.6%
4	-.05	.89	23.6%

	Father's eyes 1	2	3	4
b) FITTED COUNTS:				
(T)	358	264	180	198
1	103.8	76.6	52.3	57.4
2	113.9	84.0	57.2	63.0
3	55.9	41.2	28.1	30.9
4	84.6	62.3	42.5	46.8
c) FITTED SQUARE ROOTS: $\sqrt{4(\text{fit}) + 1}$				
1	20.4	17.5	14.5	15.2
2	21.4	18.4	15.2	15.9
3	15.0	12.9	10.6	10.2
4	15.4	16.0	13.1	13.7
d) OBSERVED SQUARE ROOTS: $\sqrt{4(\text{obs}) + 2}$				
1	27.9	16.8	12.9	11.0
2	18.3	22.4	12.9	12.1
3	10.1	11.7	14.9	9.7
4	15.1	12.1	13.2	21.0
e) DIFFERENCES: of last two				
1	7.5	-0.7	-1.6	-4.2
2	-3.1	4.0	-2.3	-3.6
3	-4.9	-1.2	4.3	-0.5
4	-0.3	-3.9	0.1	7.3

Examples: 103.8 = 29.0% of 358

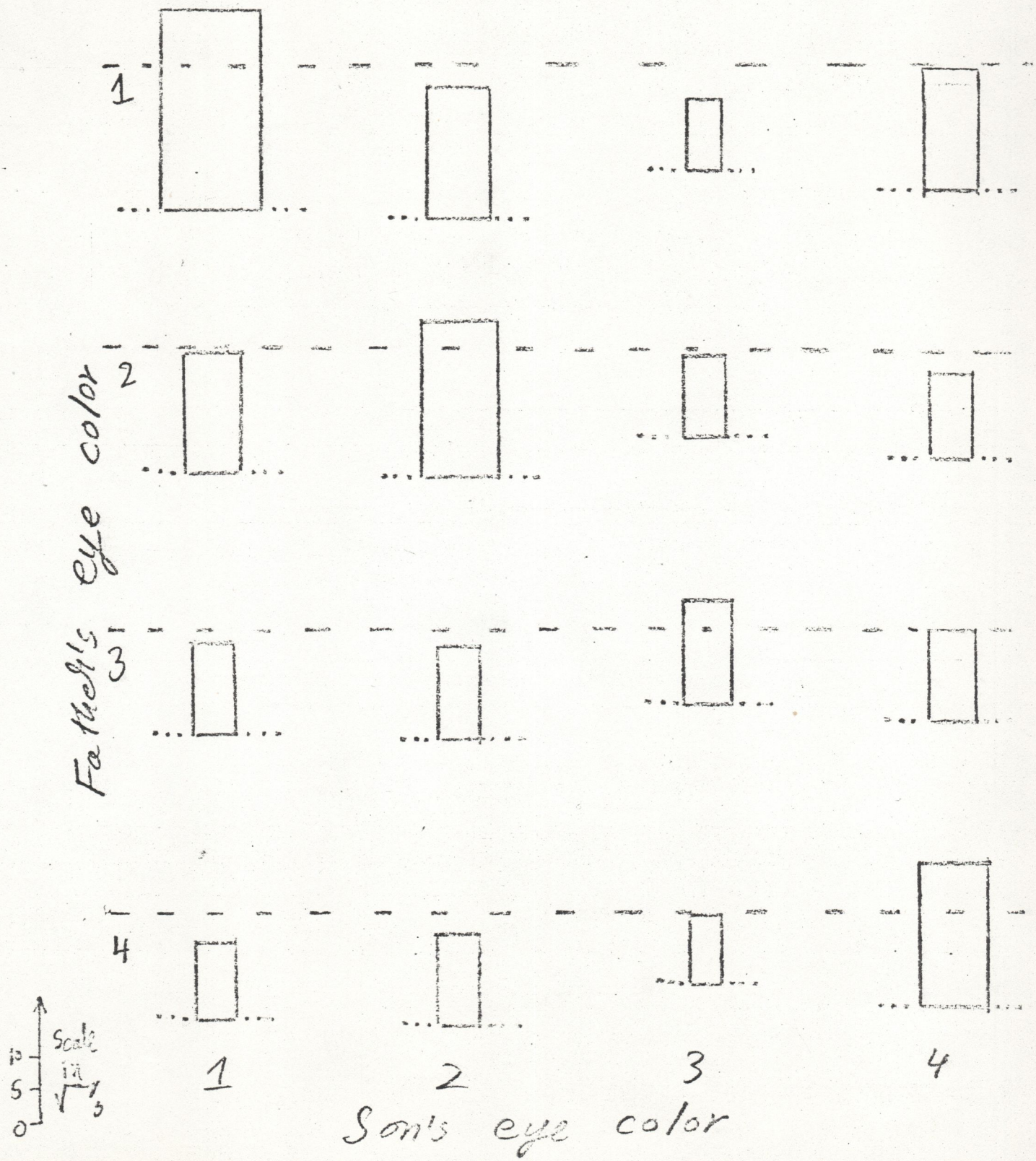
$20.4 = \sqrt{4(103.8) + 1}$

$27.9 = \sqrt{4(194) + 2}$

7.5 = 27.9 - 20.4

PLOT 5

Suspended rootobar plots of son's eye color for each father's eye color. (Allowance made for distribution of father's eye color.)



that the sons are the fathers of the next generation. But no such argument can be applied to the data in table 6, which reports blood group distributions for six tribes of Arizona indians. There is little sense in picking out an Rh pattern and asking how it is distributed over tribes of indians. There is much sense in picking a tribe of indians and asking how they are distributed over Rh patterns.

Many tables that might be thought of as contingency tables are unwisely thought of as contingency tables deserving symmetrical treatment. It seems that more tables deserve unsymmetrical treatment than deserve symmetrical treatment.

table 6 of chapter 13

Rh Patterns for Six Tribes of Arizona Indians

a) PATTERNS: reaction = +, nonreaction = -, not tested = blank

Pattern number	Rh-antiserum			
	rh′	hr′	Rh′	rh″
1	+	-	+	-
2	+	+	+	-
3	-	+	+	+
4	+	-	+	+
5	+	+	+	+
6	-	+	+	-
7	+		+	+
8	+		+	-

b) DISTRIBUTIONS:

Tribes	Patterns								(Total)
	1	2	3	4	5	6	7	8	
West Navaho	24	14	29	1	15	4	15	2	(104)
Apache San Carlos	47	22	64	2	40	3	0	0	(178)
Pima	136	59	104	20	139	6	22	3	(489)
Yuma	63	32	34	4	39	7	0	0	(179)
Hopi and Tewa	20	23	35	0	7	21	6	5	(117)
Maricopa	12	3	10	0	8	0	5	0	(38)

Source: K. S. Brown, B. L. Hanna, A. A. Dahlberg, H. H. Strandskov 1958. The distribution of blood group alleles among indians of southwest North America. Amer. J. Human Genetics 10: 175-195 (esp. Table 1 on page 177).

13C. <u>One factor ordered; unsymmetric treatment</u>.

Many "contingency tables" involve at least one "ordered factor", at least one factor -- either "rows" or "columns" or both -- whose versions are ordered. Their analysis badly needs to take this into account whenever it happens. In the present section, we will confine ourselves to the case where only one factor is ordered, leaving the cases with two ordered factors for the next section.

Such a limitation does not bring us down to a single case. Three cases need to be distinguished:

- the table should be approached unsymmetrically; the ordered factor is consequent -- we are really dealing with an ordered response to an unordered factor.
- the table should be approached unsymmetrically; the ordered factor is antecedent -- we are really dealing with an unordered response to an ordered factor.
- the table should be approached symmetrically -- our need is to take account of the order after carrying out a symmetrical analysis.

As a first example, consider the data in table 7. Brownlee (1949, page 46) reports the qualities found for many samples of products made on each of three shifts. (He notes that the raw material supplied to each shift can be taken as of equivalent quality.) Since we are here dealing with an ordered response -- quality -- masquerading as a factor, the first thing to do is to make this obvious. The first step,

table 7 of chapter 13

Brownlee's Example of Quality of Product

(a) COUNTS: Number of occurrences of specified quality

Shift	Quality a	b	c	d	e	f	(*)
A	11	23	8	5	18	18	(83)
B	17	29	10	17	7	15	(95)
C	6	21	8	24	15	9	(83)
(*)	(34)	(73)	(26)	(46)	(40)	(42)	((261))

(b) CUMULATIONS: Number of occurrences of at least specified quality

	≥a	≥b	≥c	≥d	≥e	(*)
A	11	34	42	47	65	(83)
B	17	46	56	73	80	(95)
C	6	27	35	59	74	(83)

(c) F-ROOTS: of corresponding counted fractions, multiplied by $1/\sqrt{2}$ (i.e. $\sqrt{f} - \sqrt{1-f}$, not yet changed to $\sqrt{2f} - \sqrt{2(1-f)}$)

A	-.57	-.13	.01	.10	.34
B	-.48	-.02	.13	.40	.52
C	-.69	-.25	-.11	.31	.61

(d) ELEMENTARY ANALYSIS: of table of f-roots, by midmean differences

							Hinges for residuals
A	.05	.03	.04	-.12	-.05	-.02	
B	.00	.00	.02	.04	-.01	.12	-.05
C	-.05	-.07	-.06	.11	.24	-.04	+.04
	-.63	-.17	-.04	.21	.38	.03	

(*) = Totals

Note: Presumably a is the best quality.

Examples: 11 = 11, 34 = 11 + 23

$$-.57 = \sqrt{11/83} - \sqrt{72/83}$$

Source: K. A. Brownlee 1949. Industrial Experimentation (4th ed.) (esp. table on page 46).

see panel (b) of table 7, is to find the counts that would be found if the responses were to be judged 'higher' or 'lower' at each possible cutting point. (In this data we can cut between 'a' and 'b', between 'b' and 'c', and so on to between 'f' and 'g'.) At each cutting point we have a counted fraction for each of the three shifts, and our next step is to express all these counted fractions effectively. In table 7 we have chosen to do this by taking f-roots; clearly we might have chosen to take f-logs instead. The elementary analysis comes out quite clearly:

- shift B appears to perform somewhat better than shifts A and C. (The small size of most residuals adds persuasiveness to this suggestion.)

- beyond this, shift C seems to have fewer occasions with very low quality -- quality **f**, and possibly qualities **e** and **f** together, while shift A may be light on quality **d** and possibly quality **e**. (Might this be a peculiarity of the inspector or inspectors on these shifts?)

- the difference between qualities $\geq$ **a** and $\geq$ **b** seems to be greater than for other adjacent pairs of cutting points. Thus quality **b** may extend over a wider range than qualities **c**, **d**, or **e**.

Treating this data as an ordered response -- quality -- to an unordered factor -- shift -- has led to a clean and effective analysis.

This case -- ordered response masquerading as a factor -- comes up often and in many fields. Table 8 gives a further illustration, using results of G. Hartman (given by

table 8 of chapter 13

Hartmann's Data on Tasting Phenylthiocarbamide

	Number found	
Threshold	Men	Women
10	15	42
9	35	52
8	46	38
7	31	30
6	23	19
5	13	17
4	9	6
3	7	5
2	10	10
1	13	19
0	25	33
<0	63	43
(Total)	(290)	(314)

Source: K. A. Mather 1949. Statistical Analysis in Biology. (Esp. Table 49 on page 196.)

Mather, 1947) concerning the tasting of phenylthiocarbamide by men and women. This is one of a number of substances for which the ability to taste varies widely and is inherited. Hartman used 11 strengths of solution, and called "threshold" the strongest solution that could not be tasted. (Thus <0 means that the weakest solution, 0, could be tasted.)

Conversion -- first to counted fractions and then to f-roots -- at each cutting point gives quite consistent results; the apparent difference between men and women varying irregularly between .21 and .10 (on the f-root scale) with men more sensitive. Clearly this analysis is better than any 'contingency-table' analysis.

* * *

For an example of the second case, where the single ordered "factor" is of the nature of a possible cause, while the unordered one is of the nature of a response, we may turn to counts given by Thieme (1952) for the distribution of OAB blood groups in Puerto Rico. Thieme classified all his subjects (by combining "skin color", "hair form", and "lip thickness and nose shape") on a scale from 0 (most typically white) to 9 (most typically negro) and gave blood-group distributions for each class. Only six individuals were "0" and only one individual was "9", so these two classes have been omitted from table 9.

In looking into this data we are concerned with how the distribution of blood groups shifts as we move along the ethnic scale. Since the fraction in each of the first four blood groups is relatively large, we can approach the analysis of this data by examining how each of these four fractions vary as we move along that scale. It again seems natural and

table 9 of chapter 13

Thieme's Blood Group Distributions for Eight Levels of Ethnic Appearance in Puerto Rico

a) COUNTS:

Ethnic Appear.	Blood group O	A_1	A_2	B	A_1B	A_2B	(*)
1	50	27	11	10	1	2	(101)
2	450	214	88	63	9	4	(828)
3	787	382	136	112	8	16	(1441)
4	224	109	40	50	6	4	(433)
5	70	26	9	22	5	1	(133)
6	89	28	14	24	1	3	(159)
7	54	15	10	18	1	3	(101)
8	23	4	3	10	1	1	(42)
							((3238))

b) PERCENTS: among those of given ethnic appearance

1	49.5	26.7	10.9	9.9	-	-
2	54.3	25.8	10.6	7.6	-	-
3	54.6	26.5	9.4	7.8	-	-
4	51.7	25.2	9.2	11.6	-	-
5	52.6	19.6	6.8	16.5	-	-
6	56.0	17.6	8.8	15.1	-	-
7	53.5	14.8	9.9	17.8	-	-
8	54.8	9.5	7.1	23.8	-	-

c) F-ROOTS: corresponding to percents

1	-.01	-.48	-.87	-.89	-	-
2	.09	-.50	-.88	-.97	-	-
3	.09	-.48	-.91	-.96	-	-
4	.03	-.52	-.92	-.85	-	-
5	.05	-.63	-1.00	-.72	-	-
6	.12	-.69	-.93	-.77	-	-
7	.07	-.77	-.89	-.68	-	-
8	.10	-.91	-.99	-.54	-	-

d) MIDMEANS OF END THIRDS: raw thirds.

1,2,3	.07	-.49	-.89	-.94	-	-
6,7,8	.10	-.78	-.93	-.67	-	-

e) COUNTS: for approximately one third of cases.

1,2	500	241	99	73	(10)	(6)	(929)
4,5,6,7,8	460	182	76	124	(14)	(12)	(868)

(further steps left to reader)

(*) = Total

Source: F. P. Thieme 1952. The geographic and racial distribution of ABO and Rh blood types and tasters of PTC in Puerto Rico. Amer. J. Human Genetics 4: 94-112. (Esp. Table 3 on page 101.)

reasonable to express these fractions as f-roots. Having done this, we can expose the results to view as in plot 10. We see here that:

- the frequency of Group O appears to trend slowly upward.
- that of Group A_1, after an uncertain start, trends firmly downward.
- that of Group A_2 trends slowly downward.
- that of Group B, after an initial hook, trends firmly upward.

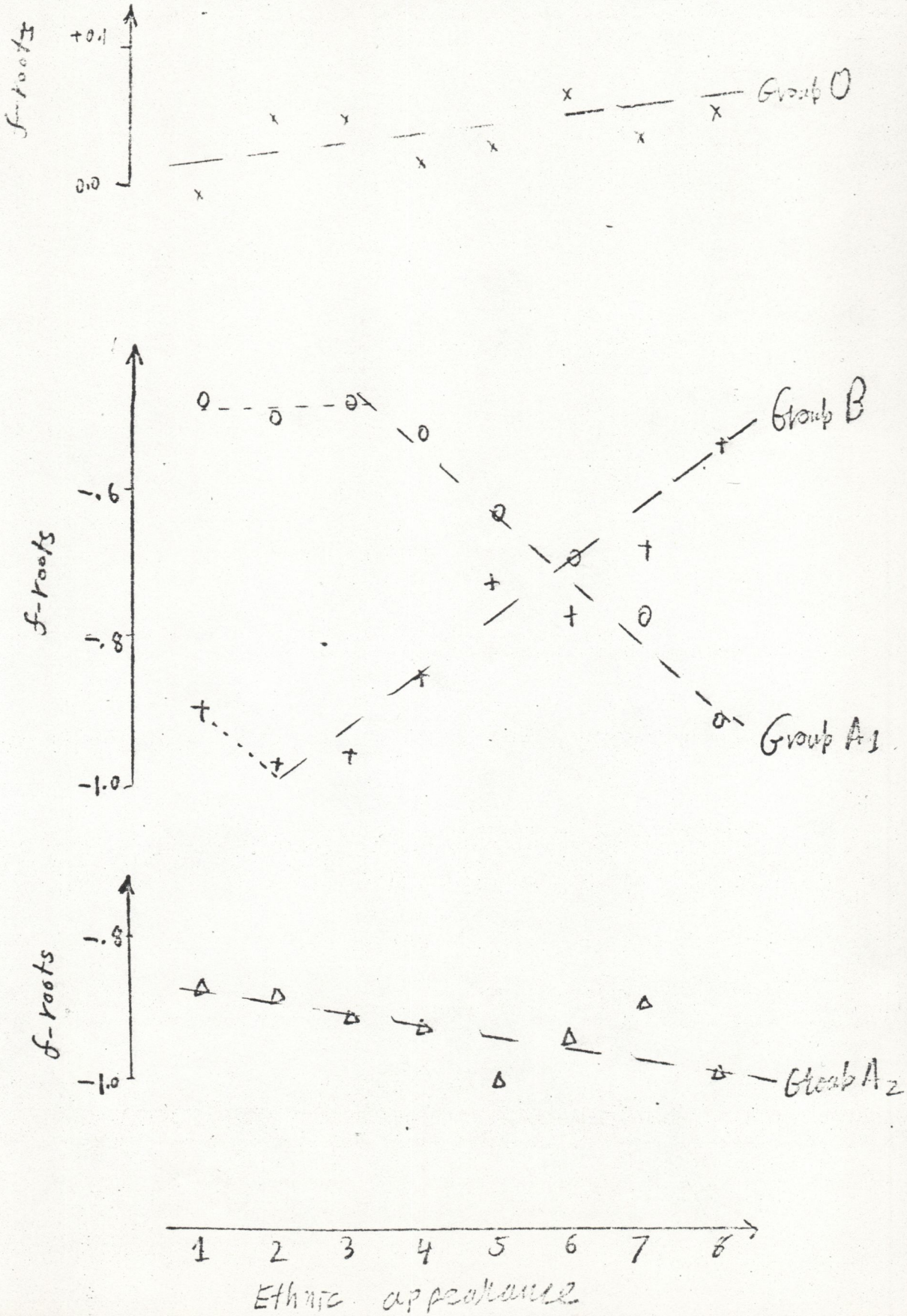
If we were foolish enough to want to avoid making a plot, we could attempt to assess some sort of trend arithmetically for each blood group. Two ways may seem fairly natural:

- that carried out in portion (d) of table 9, where we have taken the midmeans of the results for about 1/3 of the scale values at each end. (The overall indications these give are consistent with the results of the plot, though perhaps weaker. They necessarily fail to detect the "hooks" shown by the plot.)
- that begun in portion (e) of table 9, in which about 1/3 of all the individuals at each end are thrown together. (The calculation of %'s and f-roots has been left to the reader.)

What has been done in this example is reasonably typical of the best we know how to do when one factor is ordered

plot 10 of chapter 13

Dependance of blood group fractions on ethnic appearance



and the other involves fairly large classes. If the unordered factor has very many small cells, we are in much more difficulty. Sometimes a wise or ingenious way of combining cells into supercells -- possibly in more than one way -- can help us, though there are likely to be questions of subjectivity. (After all, making counted fractions out of an ordered factor operates by combining cells into supercells in several ways, so it is not unlikely that more than one way will help us here.)

13D. Symmetric factors, one ordered.

So much then for the two unsymmetric cases, what about the symmetric one? For this case, we can probably reach the needed insight by using an imaginary example. Let us consider a table of counts of adult people where the two factors are:

- eye color.
- height.

Here height is ordered, while eye color is not. Neither is antecedent to the other, so it would seem that we do not want to make an unsymmetric analysis. However, it is not clear what form of "symmetric" analysis makes very good sense.

For the present, then, we must plan to tackle such tables in a somewhat 'catch as catch can' way. Specifically, it would seem sensible to:

- apply the sort of (unsymmetrical) analysis we have just been applying to the Puerto Rico blood group data.
- go to the other extreme, and treat height as a response.

- begin symmetrically by using either marginal totals or elementary analysis of log counts to provide fitted values, obtaining residuals in square root terms, and then proceed, unsymmetrically, looking for tilt in these residuals as a function of height.

It might indeed be wise to try two -- perhaps all three -- of these approaches, selecting the one which seems to give the cleanest analysis.

13E. Both factors ordered; unsymmetric treatment.

As an entry into the problems we must face when both factors are ordered, let us begin with the data by Burgess and Cottrell (1936) on the relation between husband's education and marriage-adjustment scores shown in table 11. Clearly we are allowed to take marriage-adjustment score as a response. (Whether it is a response to education itself, to that increased ability usually accompanies greater education (perhaps as a cause, not an effect) or to the common effects of parental attitudes on desire for education and ability to adjust is not a question to be answered from this set of data.) (One reader suggested that it might be the effect of well-adjusted wives putting husbands through college.) So again we ought to accumulate counts, convert to f-roots or f-logs and proceed to an elementary analysis. Here, except for a few isolated irregularities (-.13 among Graduate Work, -.09 among Grade School) apparently concentrated in the smallest classes, the fit by stripes seems quite good. The most meaningful results are the stripe adjustments for education, namely

Graduate Work:	-.19
College:	-.08
High School:	+.11
Grade School:	+.15

table 11 of chapter 13

Counts of Combinations of Husband's Education and Marriage Adjustment Scores

a) COUNTS:

Husband's Education	Marriage adjustment score: Very low	Low	High	Very high	(T)
Graduate Work	4	9	38	54	(105)
College	20	31	55	99	(205)
High School	23	37	41	51	(152)
Grade School	11	10	11	19	(51)

b) CUMULATED COUNTS:

	≤ Very low	≤ Low	≤ High	(T)
Graduate Work	4	13	51	(105)
College	20	51	100	(205)
High School	23	60	101	(152)
Grade School	11	21	32	(51)

c) F-ROOTS: multiplied by .707 = $1/\sqrt{2}$ (again not yet changed)

Graduate Work	-.78	-.58	-.02
College	-.64	-.37	.02
High School	-.53	-.15	.24
Grade School	-.42	-.12	+.18

d) ELEMENTARY ANALYSIS: of these f-roots, by midmean* differences.

					Hinges for residuals
Graduate Work	-.01	-.10	+.05	-.20	
College	+.02	.00	-.02	-.09	
High School	-.06	+.03	+.01	+.10	-.04
Grade School	+.01	+.02	-.09	+.14	+.02
	-.29	.00	+.41	-.28	

midmean* = midmeans of 4 but medians of 3.

Examples: 4 = 4, 4 + 9 = 13.

$$-.78 = \sqrt{4/105} - \sqrt{101/105}.$$

$$-.78 = (-.01) + (-.20) + (-.29) + (-.28).$$

which show a steady shift (in the direction of better adjustment) with increasing education.

So far so good, but we will clearly be better off -- for two reasons -- if we can replace the ordered variable "husband's education" by a wisely numbered variable, namely:

- in situations where either the effect is small or many different levels of the ordered factor occur, we can make a more sensitive inquiry if we can look explicitly for tilt, rather than just looking for an ordered trend.
- in situations -- like our example -- where there is no difficulty spotting a trend, we have a better chance to ask about its character if we can plot numbers against numbers.

One way to do this -- a way that may sometimes prove nonsensical or unwise, but a way that can always be tried -- is to use the distribution of the ordered factor itself to establish a scale. Table 12 shows one way to do this in our example. We find the median and both hinges of the ranks of each level of our ordered variable, then convert each into f-roots -- or, when it serves us better, into f-logs -- and form their skeleton midmean. (In this example, the skeleton midmean exactly matches the median in three cases and differs by 0.01 in the fourth. Thus, in this case, we might as well have used only the median.) Since we are only going to use the results for plotting, a constant factor makes essentially no difference. Thus the factor of $0.707 = 1/\sqrt{2}$ introduced in table 12 is wholly unimportant.

table 12 of chapter 13

Calculation of Plotting Values for Four Levels of Husband's Education

Husband's Education	Inverse ranks* Covered	H+M*	f-roots**	Midmean***
Graduate work	1 to 105	27	-.74	
		53	-.63	-.63
		79	-.53	
College	106 to 310	157	-.26	
		208	-.14	-.13
		259	+.01	
High School	311 to 462	348.5	+.26	
		386	.38	+.38
		424.5	.49	
Grade School	463 to 513	475	+.69	
		488	.76	+.76
		501	.84	
(Total)		(513)		

Ranks* = ranks for education. H+M* = hinges and median.
f-roots** = values of (f-root)/$\sqrt{2}$ (again not yet changed).
Midmean*** = skeleton midmean.
Examples: 105 = 105, 310 = 105 + 205

$$53 = \tfrac{1}{2}(1+105),\ 27 = \tfrac{1}{2}(1+53),\ 79 = \tfrac{1}{2}(53+105)$$

$$-.74 = \sqrt{27/513} - \sqrt{486/513}$$

$$-.63 = \tfrac{1}{4}(-.74 + 2(-.63) + (-.53))$$

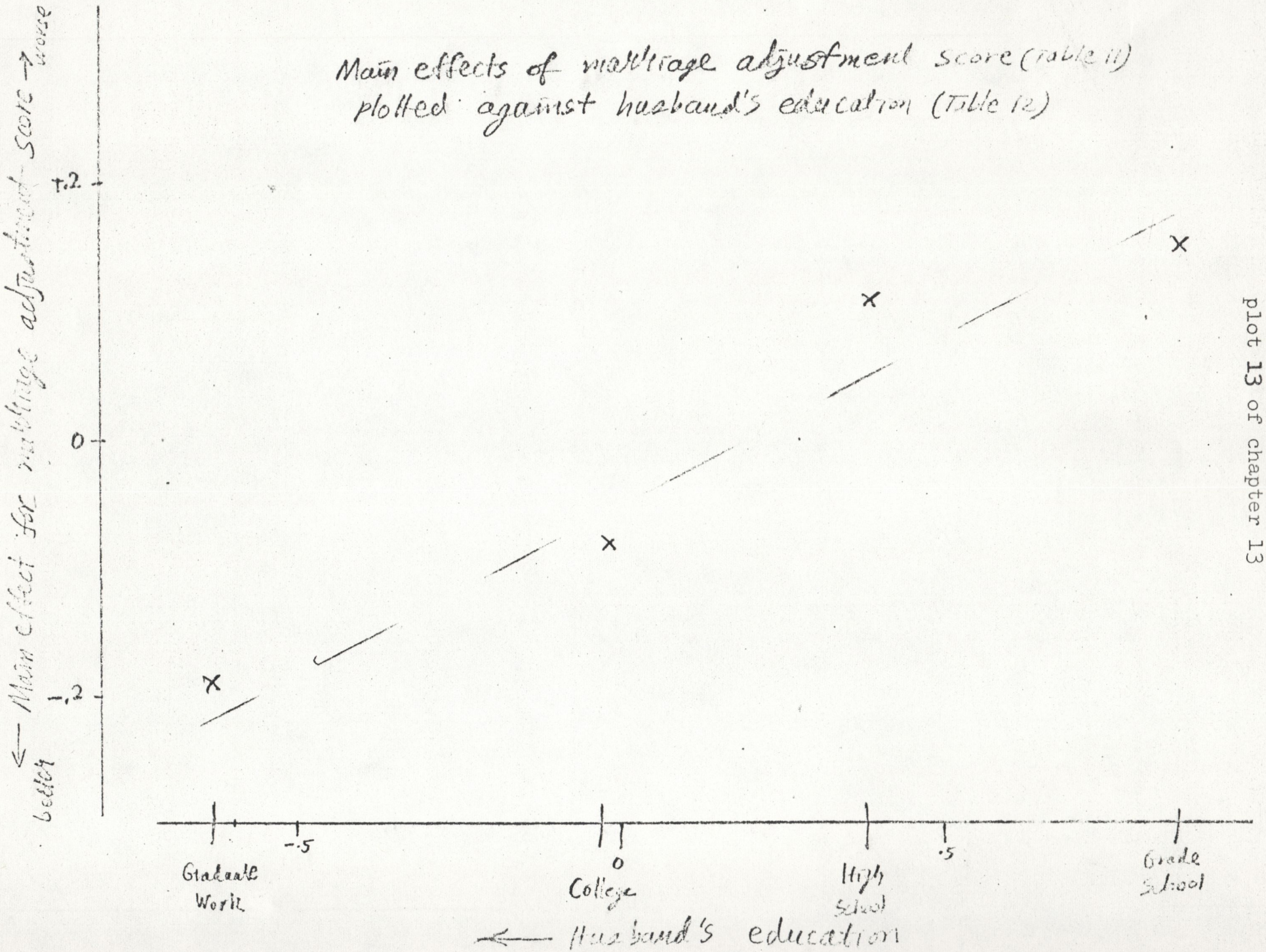
Plot 13 shows the results of plotting the main effects for husband's education found in table 11 against the plotting values found in table 12. The resulting picture suggests either:

- considerable irregularity, or
- a larger break between high school and college than our procedure of finding numbers for husband's education would forecast.

Clearly this analysis has been quite effective.

(Another analysis applicable in this case will be illustrated in section 13M toward the end of this chapter.)

plot 13 of chapter 13



13F. Both factors ordered; symmetric treatment desired.

Not every two-way table of counts clearly deserves either unsymmetrical or symmetrical treatment. Consider Gilby and Pearson's data (1911) on the distribution of school children according to 'intelligence' and 'standard of clothing' as given in table 14. (Notice that the order of intelligence runs from A = mentally deficient to G = very able, and not in the order that we might expect.)

It is clearly possible to argue that we should look at this table with 'intelligence' preceding 'standard of clothing'. By and large, more intelligent children have more intelligent parents -- more intelligent parents are less likely to be poor, and are hence more likely to have well-dressed children.

It is almost equally possible to argue that clothing should precede intelligence. The Coleman Report (Equality of Educational Opportunity, 1966) underscored the fact that, by and large, more deprived children come to do less well on aptitude tests as well as ability tests. Accordingly, less well clad children of equal inborn ability are likely to demonstrate less "intelligence" after a few years in school.

If we can argue for both unsymmetrical approaches, we must be able to argue for a symmetrical approach. This we can do, and in at least two ways:

- as a refuge from bafflement as to which of the unsymmetric approaches is to be taken
- because, as it is argued, both "intelligence" and "standard of clothing" are consequences of the social structure in which the children were conceived and grew up.

table 14 of chapter 13

Distribution of British Schoolchildren According to Intelligence and Standard of Clothing

a) SCALE OF INTELLIGENCE:

A = mentally deficient

B = slow and dull

C = dull

D = slow but intelligent

E = fairly intelligent

F = distinctly capable

G = very able

b) COUNTS:

	Intelligence						
Clothing	A and B	C	D	E	F	G	(T)
Very well clad	33	48	113	209	194	39	(636)
Well clad	41	100	202	255	138	5	(751)
Poor but passable	39	58	70	61	33	4	(265)
Very badly clad	17	13	22	10	10	1	(73)
							((1725))

(T) = Totals

Source: W. H. Gilby and K. Pearson 1911. On the significance of the teacher's appreciation of general intelligence. Biometrika 8: 93-108, especially page 94 (as quoted by M. G. Kendall (1943,...,1953) Table 13.1, page 320).

In this example, it is not clear that any one approach is right -- indeed, it almost seems clear that NO one approach is correct. Fortunately, not all tables leave us so uncertain as to what we ought to do. Consider Woo's data on laterality of hand and eye, which is set out in table 15. There seems to be no reasons to treat handedness and eyedness unsymmetrically but many to treat them quite alike.

What then are we going to do with two-way tables of counts that deserve symmetric treatment and have both factors ordered? We are, so far as the present account goes, really in no better situation than we were with one variable ordered (see Section 13D). We have the same three sorts of alternatives:

- even if it should be treated symmetrically, treat it unsymmetrically one way,
- or the other way, or
- begin symmetrically by obtaining residuals, and then use the order in both factors to examine these residuals carefully.

While these may seem counsels of desperation:

- they are about as well as we know how to do (at least without going far beyond our present scope),
- especially when two or more are tried, and the one with cleanest results chosen, they do serve us quite well.

(If we were prepared to have and use tens of pages of tables of the two-way correlated Gaussian distribution, we could set up other ways to deal with such tables, but this is far outside our present scope.)

In this example, it is not clear that any one approach is right -- indeed, it almost seems clear that NO one approach is correct. Fortunately, not all tables leave us so uncertain as to what we ought to do. Consider Woo's data on laterality of hand and eye, which is set out in table 15. There seems to be no reasons to treat handedness and eyedness unsymmetrically but many to treat them quite alike.

What then are we going to do with two-way tables of counts that deserve symmetric treatment and have both factors ordered? We are, so far as the present account goes, really in no better situation than we were with one variable ordered (see Section 13D). We have the same three sorts of alternatives:

- even if it should be treated symmetrically, treat it unsymmetrically one way,
- or the other way, or
- begin symmetrically by obtaining residuals, and then use the order in both factors to examine these residuals carefully.

While these may seem counsels of desperation:

- they are about as well as we know how to do (at least without going far beyond our present scope),
- especially when two or more are tried, and the one with cleanest results chosen, they do serve us quite well.

(If we were prepared to have and use tens of pages of tables of the two-way correlated Gaussian distribution, we could set up other ways to deal with such tables, but this is far outside our present scope.)

table 15 of chapter 13

Woo's Counts on Laterality of Hand and Eye

	Ocular laterality for general astigmatism		
Hand*	"Left-eyed"	Ambiocular	"Right-eyed"
Left-handed	34	62	28
Ambidextrous	27	28	20
Right-handed	57	105	52

*Manual laterality as measured by a balancing device.

Source: T. L. Woo 1928. Dextrality and sinistrality of hand and eye. Second Memoir. Biometrika 20A: 79-148.

(Another possibility would be to try the technique illustrated below in section 13M, whose symmetry is noted in section 13N, where the expressions used are called ff-logs. As yet, we have too little experience with this method to be clear as to how well it will work.)

13G. What to do with big cells -- and with prominent diagonals

When, in 13A, we treated two-way tables of counts symmetrically and without using any structure in either factor -- when we treated them as contingency tables -- we treated the count in each cell as if it were a pure count, not the numerator of a counted fraction. So long as all our cells are small -- so long as their counts are only small fractions of the total count -- this is obviously sound. For if we were to treat them as counted fractions, we would find the f-roots almost exactly proportional to the (doubled) square roots of the counts -- or the f-logs almost exactly proportional to the logs of the counts.

But what if there are some large cells -- some cells so large as to make working with counted fractions different in result from working with counts? Which should we use? Square roots or f-roots? Logs or f-logs?

And the echo comes back "which?". The best advice so far available is very general:

- when in doubt, try both -- and use the one that gives the cleanest analysis.

* * *

Some two-way tables of counts have the same set of labels for rows as for columns. Such tables have (main) diagonals, consisting of those cells whose row labels match their column labels.

Another issue we should face is illustrated by table 3, where the main diagonal entries appear -- at least as soon as we look at residuals -- noticeably different from the rest of the table. Having noticed as systematic an effect as this, we should certainly plan to shift it from residuals to fit. The essential step in doing this is to find a way of choosing a diagonal adjustment that:

- is transparent to a uniform shift of the whole main diagonal
- is immune to shifts either uniform by rows or uniform by columns
- is resistant to erratic or weird values.

Once we have such a step, we can use it in preparation for an elementary analysis.

Fortunately, finding such a step is not hard to do.

Let us begin by looking at the NW corner of panel **b** of table 3. The four corner values are

2.29	1.85	...
1.92	2.09	...
...	...	

with 2.29 and 2.09 on the diagonal. If we form

$$2.29 - 1.85 - 1.92 + 2.09 = 0.61$$

we have a number that is immune to changes uniform by rows or by columns -- because the pattern of + and - signs produces cancellation. This number, far from being immune to uniform changes on the diagonal, responds by exactly twice the change.

We can calculate a similar matched value for each pair of diagonal entries. We find the results shown in panel **a** of table **16**, which lead to a diagonal adjustment of +0.45. Subtracting 0.45 from all the diagonal entries leaves the values in panel **b**, whose elementary analysis is given in panel **c**. Reassembling, we have the d i e l e m e n t a r y a n a l y s i s given in panel **d**.

When we turn to the comparison of the stripe adjustments found in this analysis with that found in table 3, panel **e** shows that both row and column adjustments continue about as before. (If we had some reason to expect equal numbers of each eye color, we might argue that the dielementary row adjustments are more sharply focussed, since three of the four are closer to zero, but such a remark seems irrelevant here.

Panel **f** emphasizes the great reduction in size of the residuals in the dielementary analysis, and the absence of any tendency for them to straggle.

In this instance, the dielementary analysis has improved the analysis considerably.

table 16 of chapter 13

Extended Elementary Analysis (Dielementary Analysis)
of the Values of table 3(b)

a) THE DOUBLE DIFFERENCES:

For		
2.29 and 2.09	2.29-1.85-1.92-2.09	= .61
2.29 and 1.74	2.29-1.61-1.40-1.74	= 1.02
2.29 and 2.04		= 1.09
2.09 and 1.74		= .69
2.09 and 2.04		= 1.01
1.74 and 2.04		= .79

midmean of these six = 0.90
diagonal adjustment = $\frac{1}{2}$(0.90) = 0.45

b) THE LOGS ADJUSTED ON THE DIAGONAL:

	1	2	3	4
1	1.84	1.85	1.61	1.48
2	1.92	1.64	1.61	1.56
3	1.40	1.53	1.29	1.36
4	1.75	1.56	1.63	1.59

c) ELEMENTARY ANALYSIS OF ADJUSTED VALUES: by midmean differences

1	.03	.13	.00	-.10	.02
2	.11	-.08	.10	-.06	.02
3	-.11	.11	.08	.18	-.28
4	-.03	-.13	.05	.04	-.01
	.13	.04	-.07	-.10	1.65

d) COMPLETE DIELEMENTARY ANALYSIS: reassembling a and c

1	.03	.13	.00	-.10	.02
2	.11	-.08	.00	-.06	.02
3	-.11	.11	-.02	.08	-.28
4	-.03	-.13	.05	.04	-.01
	.13	.04	-.07	-.10	1.65

and +.45 on the diagonal

table 16 (Cont'd)

e) COMPARISON OF ELEMENTARY AND DIELEMENTARY STRIPE ADJUSTMENTS

Stripe	Analysis	1	2	3	4
Rows	Elem.	.04	.08	-.23	-.05
	Dielem.	.02	.02	-.28	-.01
Cols.	Elem.	.10	.07	-.06	-.15
	Dielem.	.13	.04	-.07	-.10

f) COMPARATIVE SIZES OF RESIDUALS:

Analysis	Typical square	Typical midsquare
Elem.	.0503	.0280
Dielem.	.0065	.0097

13H. How far have we come?

This chapter has been devoted to the care and treatment of two-way tables of counts. Care has been called for in selecting the treatment to be given. The importance of just how we are looking at a given table has been emphasized. (Probably not enough has been said about the reasonableness -- and moderately frequent wisdom -- of treating one-and-the-same table from two or more points of view.)

The most basic points we have been making have to do with ordered factors. Specifically:

- if one factor is a concealed ordered response -- if we can interpret its gradations as corresponding to a family of alternative cutting points -- we dare not treat it as just another factor, we must first cumulate and then re-express -- probably as f-roots or f-logs.

- if one factor is -- or can be reasonably taken as -- a response -- ordered or not -- while the second factor is ordered -- we dare not treat this second factor as just another factor -- we owe it to ourselves to plan to look for tilt -- or even bend or twist -- in the way the response(s) depend on this ordered factor.

Under either or both of the circumstances just described, our analysis is inevitably pulled far away from the analyses set out in most texts as what is appropriate for a contingency table. Either we cumulate and re-express to change our table into a two-way table of responses, or we start more conventionally but relapse to the sort of analysis

that would be appropriate when looking for tilts, etc. in a two-way table of responses. In one way or another, either of these situations moves us back toward the analyses appropriate for responses. The special characteristics of counts are only reflected in what we do to make the reductions.

When neither of these constraints holds, but the relation of one factor to the other is logically more or less that of 'before' to 'after', we still need to analyze the data in an unsymmetric way, focusing on the various distributions over the 'after' factor that arises, one for each version of the 'before' factor. (See section 13B for further discussion and plot 5 for an illustration.)

There remain those cases that are truly worthy of the name contingency tables, where neither factor has a 'before' to 'after' relation to the other -- where it is natural to make an analysis that treats the two factors symmetrically. It is here that we have failed to offer as much help as we would have hoped.

If neither factor is usefully structured -- in particular if no factor is ordered -- we can do little else than treat our table as a contingency table in the classical sense, using any of at least three analyses:

- using marginal totals to set reference values and basing residuals on differences of (doubled) square roots (as in table 1 above).

- taking logs, using their midmean differences to set reference values, and basing residuals on differences of logs (as in table 2 above).

- taking logs and midmean differences to set reference values, but then basing residuals on (doubled) square roots. (These residuals can now be analyzed in any way we know. In particular, if one or both factors are ordered we can look for trend by any reasonable method.)

When we have no 'before' to 'after' relation, but one or both factors are ordered, the situation is more complicated. It has, however, the complications of opportunity, not those of blind despair. Such orderings challenge us to make adequate use of them in our analysis, something that -- as yet -- we are unable to do without exception.

In such situations, where one ordered factor -- or two such -- exist but neither factor is a concealed response to the other, there is often not much to do except try a variety of approaches -- including 'before' to 'after' in both directions as well as organized schemes for squeezing information out of the two-way table of residuals -- and then settling upon the choice that gives the cleanest analysis.

We have also seen how to treat tables where row and column adjustments need to be supplemented with diagonal adjustments.

13K. Easy f-roots and f-logs with the slide rule.

Tables of f-roots and f-logs are not always available. In any event they require conversion of counts into fractions or % before use. Accordingly there can be advantages to more direct calculation, especially when a slide rule is available. This is particularly so when the f-roots or f-logs to be found come in sets with the same total, something that will, in particular, occur whenever we are dealing with several alternate cutting points for a response.

Suppose, then, that, for some cutting point, our counts are a and b so that

$$f = \frac{a}{a+b}$$

$$1-f = \frac{b}{a+b}$$

$$\text{f-root} = \sqrt{2f} - \sqrt{2(1-f)}$$

$$= \sqrt{\frac{a}{(a+b)/2}} - \sqrt{\frac{b}{(a+b)/2}}$$

$$\text{or } \sqrt{2a/(a+b)} - \sqrt{2b/(a+b)}$$

Thus if $a = 43$, $b = 471$, $a+b = 514$, $(a+b)/2 = 257$ the corresponding f-root is

$$\sqrt{\frac{43}{257}} - \sqrt{\frac{471}{257}}$$

One setting of the slide rule (with many settings of the cursor) suffices to read out all the numbers needed for all calculations with the same total. (On an elementary slide rule,

one resetting will be needed to overlap the other end of the scale.)

After we have set 257 on B against 1 on DF, against 43 on B we read 408 on DF and against 471 on B we read 1.353 on DF. Since

$$.408 = \sqrt{\frac{43}{257}}$$

$$1.353 = \sqrt{\frac{471}{257}}$$

we have

$$\text{f-root} = .408 - 1.353 = -.945$$

With three or more f-roots to compute with the same total, this approach should be faster than using a table to convert %s into f-roots, the %s having been found with the slide rule.

For f-logs, we write

$$\text{f-log} = 1.15\left(\log_{10}\frac{a}{a+b} - \log_{10}\frac{b}{a+b}\right)$$

$$= 1.15(\log_{10}a - \log_{10}b)$$

and calculate according to the last line. For our example we have

$\log 43 = 1.634$

$\log 471 = 2.673$

$\text{f-log} = 1.15(1.634 - 2.673) = 1.15(-1.039) = -1.195$

(The multiplications by 1.15 may either all be postponed to the end of the calculation -- and then be done together -- or (if D and L scales are both available on the body of rule) the rule can be left set for multiplication by 1.15.

This procedure, too, is likely to be easier than using a table.

13L. Sequenced factors may not be ordered

It is too easy to feel that if the versions of a factor are arranged in an order, or even naturally numbered, that the factor deserves to be treated as at least an ordered factor. This need not be the case. Factors whose versions are dates provide many examples of numbered factors that do not deserve to be considered as ordered.

Why should this be so? Because deserving to be treated as ordered means to be such that their likely effects are simple and trend-like. In many situations this is not likely to be the case for factors whose versions are dates.

Hunt (1948) gave the classification of causes of rejection of metal castings for three weeks of production. Table 17 gives the data and its contingency analysis by square-roots. The $\sqrt{}$-residuals largest in size are:

$\sqrt{}$-residual:	-5.1	+2.5	-2.5	+2.2
cause:	shift	shift	other	corebreak
week:	3	1	1	3
raw residual:	-9.4	+8.6	-7.3	+12.2

table 17 of chapter 13

Causes of Rejection of Metal Castings for Three Weeks of Production

Cause	Period 1st Wk	2nd Wk	3rd Wk	(T)
a) COUNTS:				
Sand	97	120	82	(299)
Corebreak	23	21	38	(82)
Broken	21	17	25	(63)
Drop	8	13	12	(33)
Shift	18	12	0	(30)
Misrun	8	15	4	(27)
Other	5	15	19	(39)
(T)	(180)	(213)	(180)	((573))
b) FITS: based on margins (multiplicative)				
Sand	93.9	111.2	93.9	(299)
Corebreak	25.8	30.4	25.8	(82)
Broken	19.8	23.4	19.8	(63)
Drop	10.4	12.2	10.4	(33)
Shift	9.4	11.2	9.4	(30)
Misrun	8.5	10.0	8.5	(27)
Other	12.3	14.5	12.2	(39)
(T)	(180.1)	(212.9)	(180.0)	((573))
c) SQUARE ROOTS OBSERVED: $\sqrt{\text{count}} + \sqrt{1+\text{count}} \approx \sqrt{4(\text{obsd}) + 2}$				
Sand	19.8	22.0	18.2	
Corebreak	9.7	9.3	12.4	
Broken	9.3	8.4	10.1	
Drop	5.8	7.4	7.1	
Shift	8.6	7.1	1.0	
Misrun	5.8	7.9	4.2	
Other	4.7	7.9	8.8	
d) SQUARE ROOTS FITTED: $\sqrt{4(\text{fit}) + 1}$				
Sand	19.4	21.2	19.4	
Corebreak	10.2	11.1	10.2	
Broken	9.0	9.7	9.0	
Drop	6.5	7.1	6.5	
Shift	6.1	6.8	6.1	
Misrun	5.9	6.4	5.9	
Other	7.2	7.7	7.1	

(Cont'd)

table 17 (Cont'd)

e) RESIDUALS: (c) minus (d).

Cause	1st Wk	2nd Wk	3rd Wk
Sand	+0.4	+0.8	-1.2
Corebreak	-0.5	-1.8	+2.2
Broken	+0.3	-1.3	+1.1
Drop	-0.7	+0.3	+0.6
Shift	+2.5	+0.3	-5.1
Misrun	-0.1	+1.5	-1.7
Other	-2.5	+0.2	+1.7

Source: G. A. Hunt 1948. A training program becomes a 'clinic'. Industr. Qual. Contr. (January) p. 26, as quoted by A. J. Duncan 1952. Quality Control and Industrial Statistics.

Examples:

$$93.9 = (299)(180)/573$$

$$19.8 = \sqrt{97} + \sqrt{98} = \sqrt{4(97) + 2}$$

$$19.4 = \sqrt{4(93.9) + 1}$$

$$+0.4 = 19.8 - 19.4$$

Clearly something happened in the third week. What happened is less clear. Did rejections for 'shift' become rejections for 'corebreak'? If so, was this a change in production or inspection? Or did something else happen?

In data of this sort it will often be wrong to treat week of production as ordered. Changes are frequently more likely to occur suddenly and surprisingly than they are to occur steadily and smoothly, particularly when the process is being studied intensively as the data are being collected.

Let us look next at a very different kind of data, counts of children born extramatrimonially in the district of Malmöhus Län, Sweden by five-year intervals of date and of age of mother as reported by Böök and Rayner (1950). We might argue that both date and mother's age deserve treatment as ordered factors -- clearly each could be so treated. Table 18 sets out the counts and a logarithmic contingency analysis. Once the elementary analysis is made, the residuals (panel (c)) reveal clear time-trends. Once these are also removed, the next residuals (panel (e)) reveal age-trends. Once these are also removed and minor adjustments to the date-group fits made, panel (h) shows all very small residuals with one exception: ages "30 up" in 1931-35, about whose causes we shall not speculate here.

In this situation, it was clearly wise to treat both factors as ordered, even if one was a date.

TABLE 18

table 18 of chapter 13

Age of Mother for All Children Born Extramatrimonially in Malmöhus Län, Sweden by Five-year Intervals (According to Official Records)

Mother's age	1916-20	1921-25	1926-30	1931-35	1936-40	1941-45	(***)
a) COUNTS:							
-20	1484	1694	1866	1655	1602	1494	***
20-25	3511	3494	3390	2791	2258	2053	***
25-30	1660	1474	1300	1040	821	829	***
30-35	817	655	509	409	409	435	***
35-40	455	394	279	217	213	256	***
40-45	162	152	112	85	70	86	***
45-50	11	13	9	4	4	4	***
(30-)	(1445)	(1214)	(909)	(715)	(696)	(781)	
b) LOGS: $\log_{10}$ of counts							
-20	3.17_1	3.22_9	3.27_1	3.21_9	3.20_5	3.17_4	***
20-25	3.55_3	3.54_3	3.53_0	3.44_6	3.35_4	3.31_2	***
25-30	3.22_0	3.16_8	3.11_4	3.01_7	2.91_4	2.91_9	***
30-	3.16_0	3.08_4	2.95_9	2.85_4	2.84_3	2.89_3	***
c) ELEMENTARY ANALYSIS: using compromise procedures							
-20	-.18	-.10	.00	.01	.09	.05	.08
20-25	-.03	-.02	.03	.01	.01	-.04	.31
25-30	.04	.01	+.01	-.02	-.03	-.03	-.09
30-	.07	.01	-.05	-.10	-.01	+.03	-.18
	.12	.10	.04	-.02	-.12	-.11	3.15
d) FITTED TILTS BY DATE: separately by age-group							
-20	-.13	-.08	-.03	+.03	+.08	+.13	[+.026]
20-25	-.01	.00	.00	.00	.00	+.01	[+.001]
25-30	+.04	+.02	+.01	-.01	-.02	-.04	[-.007]
30-	+.02	+.01	.00	.00	-.01	-.02	[-.004]
e) RESIDUALS AFTER DATE-TILTS:							
-20	-.05	-.02	+.03	-.02	+.01	-.08	
20-25	-.02	-.02	+.03	+.01	+.01	-.05	
25-30	.00	-.01	.00	-.01	-.01	+.01	
30-	+.05	.00	-.05	-.10	.00	+.05	

(Cont'd)

table 18 (Cont'd)

Mother's age	1916-20	1921-25	1926-30	1931-35	1936-40	1941-45
f) FITTED TILTS BY AGE:						
-20	-.05	-.01	+.04	.00	.00	-.06
20-25	-.02	.00	.01	.00	.00	-.02
25-30	.02	.00	-.01	.00	.00	+.02
30-	.05	+.01	-.04	.00	.00	+.06
	[+.017]	[+.003]	[-.013]	[.00]	[.00]	[+.022]
g) RESIDUALS AFTER BOTH TILTS						
-20	.00	-.01	-.01	-.02	+.01	-.02
20-25	.00	-.02	+.02	+.01	+.01	-.03
25-30	-.02	-.01	+.01	-.01	-.01	-.01
30-	.00	-.01	-.01	-.10	.00	-.01
(*)	(.00)	(-.01)	(.00)	(-.01)	(+.00)	(-.02)
h) RESIDUALS AFTER FURTHER ADJUSTMENT OF DATE-GROUP FITS						
-20	.00	.00	-.01	-.01	+.01	.00
20-25	.00	-.01	+.02	+.02	+.01	-.01
25-30	-.02	.00	.01	.00	-.01	+.01
30-	.00	.00	-.01	-.09	.00	+.01

[ ]: These are coefficients of simple linear functions (-5,-3,-1,+1,+3, +5 for date-groups; -3,-1,+1,+3 for age-groups).

(*) Typical values.

Source: I. A. Böök and S. Rayner 1950. A clinical and genetical study of anencephaly. Amer. J. Human Genetics 2: 61-84. (Esp. Table 9 on page 73.)

Examples:

$$3.171 = \log_{10} 1484$$

$$3.17 = (-.18) + .05 + .11 + 3.19$$

$$-.13 = (-5)(+.026)$$

$$-.05 = (-.18) - (-.13)$$

$$-.05 = (-3)(+.017)$$

$$.00 = (-.05) - (-.05)$$

$$.00 = (.00) - (.00)$$

Remark: Age group 30- used to make sizes of counts comparable. Use of 30-35 and 35- might be more effective.

This example deserves to be carried on a little further, as for instance, is done in table 19. After this further examination, a reasonable summary of the investigation of this table would be:

- the logs of counts decline steadily from date-group to date-group, any irregular additional changes are quite small.
- a strong maximum occurs in the 20-25 age-group; the shape of the fall-off to higher ages has not yet been simply described (but see remark in table 18).
- the earliest and latest five-year date-groups show additional dependence on age, with older ages increased.
- the youngest age-group shows an additional slope with date-group, higher for later groups.
- one particular value appears unusual, as already noted above.
- the remaining residuals are small.

table 19 of chapter 13

Carrying-on of the Analysis of table 17

	1916-20	1921-25	1926-30	1931-35	1936-40	1941-45
a) FITTED YEAR-GROUP ADJUSTMENTS:						
Original -- 17(c)	.11	.09	.03	-.03	-.13	-.12
Additional -- 17(g)	.00	-.01	.00	-.01	.00	-.02
Total	.11	.08	.03	-.04	-.13	-.14
* * *						
Fitted tilt	.14	.09	.03	-.03	-.09	-.14
Residual	-.03	-.01	.00	.01	.04	+.00
b) FITTED TILTS ON AGE: coefficients by date-group						
Fitted -- 17(f)	.017	.003	-.013	.000	.000	+.022

c) FITTED AGE-GROUP ADJUSTMENTS:

	-20	20-25	25-30	30-
Fitted -- 17(c)	.05	.28	-.12	-.21
* * *				
Fitted slant	+.25	+.08	-.08	-.25
Residual	-.20	+.20	-.04	+.04

d) FITTED TILTS ON DATE: coefficients by age-group

Fitted - 17(d)	+.026	+.001	-.007	-.004

13M. The treatment of mobility data

Data on the relative freedom with which individuals change, usually reported in two-way tables, represents a special case. The outstanding examples are data relating son's occupation to father's occupation. When we can tackle the questions that face us by working with a classification of occupations into groups that all will agree are ordered in a simple way, we will usually gain by treating both factors (which usually differ only in that one refers to father's, the other to son's occupation) as ordered. This means that when we go first from cell counts to cumulative counts and then bring in f-roots, f-logs, or some other way of expressing counted fractions -- we need to do this in both directions at once. This is not difficult to do, but it is different from everything we have done so far.

As an example we will use data by Rogoff (1953) on the relationship of father's and son's occupations in a sample of about 10,000 sons observed in Indianapolis in 1940. For our purposes, we shall condense the original classification somewhat, and will omit "protective service", "personal service" and "farming" as occupational categories about whose relative position there might well be dissent.

The classification that we shall use is given in panel (a) of table 20, with the relatively raw data in panel (b) and cumulations in panels (c) and (d). Let us begin by fixing our attention -- BOTH for fathers AND for sons -- on cutting point β which divides 'proprietors, managers, and officials' from 'clerical and sales'. If we consider only

table 20 of chapter 13

Analysis of a Condensed Version of Rogoff's 1940 Data
- First Phases (Completed in Table 20 or Table 21)

a) CLASSIFICATION:

Cut label	Used here	Includes from Rogoff
α	PROF	Professional Semi-professional
β	PROP	Proprietors, managers, and officials
γ	CLER	Clerical and sales
δ	SKIL	Skilled Semi-skilled
	UNSK	Unskilled

b) CELL COUNTS: cutting points are labelled, not classes

Father's	Son's occupation α	β	γ	δ
α	204	40	152	159	15
β	133	212	368	410	30
γ	141	83	461	344	26
δ	237	180	782	2550	235
	28	20	94	327	206

(Cont'd)

table 20 (Cont'd)

c) CUMULATED COUNTS:

Father's above	Son's occupation above α	β	γ	δ	(T)
α	204	244	396	555	(570)
β	337	589	1109	1678	(1723)
γ	478	813	1794	2707	(2778)
δ	715	1230	2993	6456	(6762)
(T)	(743)	(1278)	(3135)	(6925)	((7437))

d) INVERTED CUMULATIONS:

Father's below	Son's occupation above α	β	γ	δ	(T)
α	539	1034	2739	6370	(6867)
β	406	689	2026	5247	(5714)
γ	265	465	1341	4218	(4659)
δ	28	48	142	469	(675)

e) %s FOR CUMULATED COUNTS:

Father's above	Son's occupation above α	β	γ	δ
α	35.8%	42.8%	69.5%	97.1%
β	19.5	34.1	64.1	97.0
γ	17.2	29.2	64.5	97.2
δ	10.6	18.3	44.3	94.6
Father's below				
α	7.9%	15.1%	40.0%	93.0%
β	7.1	12.1	35.6	92.0
γ	5.7	10.0	28.8	91.0
δ	4.2	7.1	21.0	69.5

Source: Natalie Rogoff 1953. Recent Trends in Occupational Mobility. Glencoe; Free Press. (Esp. Table 2 on page 45.)

Examples: 204 = 204, 244 = 204 + 40, 396 = 244 + 152,...
(133 = 133, 345 = 212 + 133, 713 = 368 + 345,...)
337 = 133 + 204, 589 = 345 + 244, 1109 = 713 + 396,...
539 = 743 - 204, 1034 = 1278 - 244,...
35.8% = 204/570, 42.8% = 244/570,...
7.9% = 539/6867, 15.1% = 1034/6867,...

this one way of cutting, the available counts are boiled down to a simple 2x2 table, namely

	Son above β	Son below β	(T)
Father above β	589	1134	(1723)
Father below β	689	5025	(5714)
(T)	(1278)	(6159)	((7437))

Let us now look at this table in terms of the barrier or hurdle involved in entering the proprietorial or professional classes, the strength of the barrier at β. For father's occupation above β, the fraction of sons with occupation above β is 589/1723 = 34.1%. For father's occupation below β, the corresponding value is 689/5714 = 12.1%. (Since fathers come before sons, we should make this type of unsymmetric approach to our 2x2 table.)

If there had been 'perfect mixing' -- if father's occupation had no relation to son's occupation -- these two %s would have been the same. Clearly they differ. Some way of expressing this difference should give us a useful way of expressing the strength of the barrier at β. We are free to try any expression that we like. Whether or not we have chosen well is then a matter of to what extent the results in one table -- or in several tables -- behave simply.

Two things we are surely going to look at are f-roots and f-logs. f-logs give

For fathers above β: -.33

For fathers below β: -1.00

Deficit (f-logs): .67

while f-roots give

For fathers above β: -.32

For fathers below β: -.84

Deficit (f-roots): .52

So long as we are consistent, we can use either of these deficits -- or a deficit found in some other consistent way -- as a measure of the strength of the barrier.

In this first example, we chose to work with a barrier defined by the same cutting point for fathers and for sons. There is no reason why we should restrict ourselves so narrowly. Consider the result of taking cutting point γ for fathers and cutting point α for sons. The 2x2 table is:

	Son above α	Son below α	(T)
Father above γ	478	2300	(2778)
Father below γ	265	4394	(4659)
(T)	(743)	(6694)	((7437))

The % of sons above α are 17.2 and 5.7, and the corresponding deficits are easily computed.

Table 20, as we have seen, carries through the analysis as far as the %, leaving the f-root continuation for table 21 and the f-log continuation for table 22. Both tables of deficits are not unreasonable. Choice between them could, so far as we have now gone, depend upon our purposes, on the basis for choice. One might need to investigate many more tables before one could be clear which is to be preferred.

table 21 of chapter 13

Continuation of table 20 in terms of f-roots

	Son's occupation above				
	α	β	γ	δ	
a) F-ROOTS: for the % in table 20 (e)					
Fathers above					
α	-.28	-.14	.40	1.16	
β	-.64	-.32	.28	1.15	
γ	-.70	-.43	.29	1.17	
δ	-.88	-.67	-.11	1.04	
Fathers below					
α	-.96	-.77	-.20	.99	
β	-.99	-.84	-.29	.96	
γ	-1.03	-.89	-.43	.92	
δ	-1.09	-.99	-.61	.40	
b) DEFICITS (F-ROOTS): differences in f-roots					
					(¯)
α	.68	.63	.60	.17	(.52)
β	.35	.52	.57	.19	(.41)
γ	.33	.46	.72	.25	(.43)
δ	.21	.33	.50	.64	(.42)
(¯)	(.39)	(.48)	(.60)	(.31)	((.44))
c) CODED DEFICITS:					
α	··H	··H	···	L··	
β	···	···	···	L··	
γ	···	···	··H	L··	
δ	L··	···	···	··H	

(¯) = Means

Source: Table 20 (e).

Examples: -.28 = f-root for 35.8%
-.96 = f-root for 7.9%
.68 = (-.28) - (-.96)

Hinges for (b) are .290 and .615 (next cuts at .128, .452, and .777).

table 22 of chapter 13

Continuation of table 20 in Terms of f-logs

	Son's occupation above				
	α	β	γ	δ	
a) F-LOGS: for the % in table 20 (e)					
Father's above					
α	-.29	-.14	.41	1.75	
β	-.70	-.33	.29	1.74	
γ	-.78	-.45	.30	1.77	
δ	-1.07	-.75	-.12	1.44	
Father's below					
α	-1.22	-.86	-.20	1.29	
β	-1.28	-1.00	-.30	1.22	
γ	-1.40	-1.10	-.45	1.16	
δ	-1.57	-1.29	-.66	.41	
b) DEFICITS (F-LOGS): differences in f-logs					(*)
α	.93	.72	.61	.46	(.68)
β	.58	.67	.59	.52	(.59)
γ	.62	.65	.75	.61	(.66)
δ	.50	.54	.54	1.03	(.66)
(*)	(.66)	(.64)	(.62)	(.66)	(.65)
c) CODED DEFICITS:					
α	HIGH	..H	...	LO.	
β	...	...	...	L..	
γ	...	...	..H	...	
δ	...	...	...	HIGH	

(*) = Means.

Source: Table 20 (e).

Examples: -.29 = f-log for 35.8%
-1.22 = f-log for 7.9%
.93 = (-.29) - (-1.22)

Hinges for (b) are .54 and .695 (other cuts at .462, .618, .773, .850, .928).

Either form of deficit clearly pulls things together so that oddities in the structure of any one mobility table, or differences between any two, can be easily examined.

13N. Symmetrical analyses with both factors ordered

The analyses of the last section are unsymmetrical in approach -- we started with fathers and went to sons. When deficits are calculated from pluralities or f-roots, the analyses are unsymmetrical in effect.

When deficits are calculated from f-logs, this is not the case.

Let the counts in the four cells of a four-fold (two-by-two) table be

a	b
c	d

Then, if rows correspond to the antecedent variable, the f-logs involved are

$$1.15\left(\log_{10}\frac{a}{a+b} - \log_{10}\frac{b}{a+b}\right) = 1.15(\log_{10}a-\log_{10}b)$$

and

$$1.15\left(\log_{10}\frac{c}{c+d} - \log_{10}\frac{d}{c+d}\right) = 1.15(\log_{10}c-\log_{10}d)$$

so that the deficit is

$$1.15[(\log_{10}a-\log_{10}b) - (\log_{10}c-\log_{10}d)]$$
$$= 1.15(\log_{10}a-\log_{10}b-\log_{10}c+\log_{10}d)$$

which treats rows and columns in just the same way.

As a "doubly folded" or "four-fold" log, it is natural to call such expressions f f - l o g s .

(Formally, there is no reason for not using things that could be called f f - r o o t s . One natural way to define such a thing would be to use

$$\sqrt{4a+1} - \sqrt{4b+1} - \sqrt{4c+1} + \sqrt{4d+1}$$

In the absence of any reason why we might expect such an expression to be useful, we shall not go further in this direction.)

130. Why does the kind of factor matter?

We have recurrently come back to the question of how a particular two-way table seems most likely to be better treated. In these discussions two questions have come up repeatedly:

- is this factor really a disguised response -- do its classes correspond to having -- or being -- more or less of something which is consequent to the other factor.

- does this factor deserve to be treated as an ordered antecedent factor.

- is this factor antecedent to, consequent on, or deserving to be treated simultaneously with the other factor.

In one way or another, the answers to each of these questions have had important effects on we have chosen to analyze two-way tables of counts. Why? Are there general morals we can draw?

The question of concealed responses is probably the simplest of all. If our data has been thrown into a two-way table -- or a more complex table -- just to take care of a response that can be judged at several thresholds, we should be grateful to the careful observer or collector who has used more than one threshold in organizing the data. But we should hasten to convert a 'stripe' of such counts, first into a corresponding 'stripe' of counted fractions, and then into a better expression (f-roots? f-logs?). We can hope that the different thresholds will, once the counted fraction is expressed in a well-chosen way, behave quite similarly, thus allowing us, for most purposes, to analyze a table with one fewer factor. If we are balked in this hope, we can reasonably expect to find 'interactions' between the choice of threshold and another factor -- or other factors -- that will be interesting in and of themselves. We have every reason to convert almost every concealed response into an overt response.

Almost all -- if not all -- such situations begin with a multiway table of counts, since a 'stripe' of counts seems to be the only naive way to use several thresholds at once.

A question that arises frequently in this context -- one to which we can give no really general answer -- is this: If I have a 2-class consequent factor, shall I treat it as a response? Almost every division into two classes can be regarded as a division into 'more' or 'less'. Once we have done this, we are well on our way to treat a 2-class consequent

factor as a response. Should we go the whole way? When? When not? Answers are sometimes to be had by thinking hard about the area to which the data applies, and about the particular factors in the data. Answers are sometimes to be found by trying both approaches. The wise analyst frequently combines thinking with experimental trials of two or more approaches. Experience can -- and usually will -- help, but thought helps too.

The analysis of mobility tables, just discussed, is in some sense a situation with two concealed responses. The approach we have just taken (in 13M) can be tried on any two-way table of counts, both of whose factors can be thought of as concealed responses. It should be tried on many, probably on most.

The question of antecedent and consequent is usually central to the symmetry -- or lack of it -- of our approach. To treat all two-way tables of counts symmetrically would be to miss many, many points. Even in mobility analysis where both factors are concealed responses -- where both factors are usually based on the same classification -- it is natural -- and probably important -- to be quite unsymmetric in approach. Fathers come before sons, father's choice of occupation clearly has a much better chance to influence son's choices than vice versa. We cannot be sure that we should treat such data symmetrically.

Finally, a few more words about ordered factors. In two-way tables of counts, ordered factors come up in three quite different ways:

- when a factor is a concealed response, it is ordered.

- an antecedent factor may be ordered (shall we treat it as such?).

- one or both of a pair of 'simultaneous' factors may be ordered.

In the first case, which includes most ordered consequent factors, we are almost sure to convert the factor into a response (based upon cumulative counted fractions).

In the second case, the question is not whether the factor can be ordered but whether or not it pays to treat it as an ordered factor. The fine line between paying and not paying comes between:

- the factor that is more likely to show steady trend-like effects than more sudden and surprising changes, though both are possible.

- the factor that is less likely -- rather than more -- to show the steady effects though both kinds of effect are still possible.

The decision as to which side of this fine line a given factor stands is often difficult -- though if it is close enough to the line, either will do about as well, so who cares which choice we make? In doubtful cases, of course, trying both is likely to be called for. The basis of choice, however, is clear and well-established -- whether or not we can use it to get a clear answer in a specific instance.

13P. IEDA processor features

In its elementary version, the IEDA processor has no special facilities for square-root contingency analysis.

Since logarithmic contingency analysis is just 2-way elementary analysis of logarithms it can be easily carried out by IEDA.

The bulk of the f-root and f-log analyses considered in this chapter can be carried out within IEDA. (The appearance of f-root and f-log as one-step transformations would make their use even simpler.)

Techniques -- such as those required (i) to make overt a concealed response or (ii) to do a mobility analysis -- that depend on converting from all counts to cumulative counts are not easily applied within the present version of IEDA.

*****

13X. Example - further analysis of the mobility data

We did not push our analysis -- see tables 20, 21, 22 -- of the mobility data anywhere near an end. We were content, in tables 21 and 22, to set out two-way tables of deficits, whose 16 entries looked more or less similar, and stop there. We can certainly make elementary analyses of these tables of deficits. When we stop to look, we see that we should go further and make dielementary analyses.

While we are at this task, we may as well do the corresponding calculation for the deficits expressed in pluralities. Table 23 summarizes the situation.

The advantages of using matched expressions are shown in two ways in table 23:

- the diagonal adjustment is surprisingly unaffected by the choice of expression.
- the typical midsquare of the residuals is almost as little affected.

In addition to these numerical constancies, the far-out values for f-roots and for f-logs are in the same three cells, while the one far-out value for pluralities is in one of the same three cells.

What then changes as we change expression? In this particular example, the stripe adjustments, particularly the column adjustments, associated with cutting points for the son's occupation change notably. For f-logs, all stripe adjustments are small -- small enough to neglect in either a rough or moderately careful summary. Thus the f-log analysis is appreciably simpler. Presumably it is also more useful.

table 23 of chapter 13

Dielementary Analyses of Deficits for the Mobility Data (see tables 20,21,22)

Father's	Son's occupation above α	β	γ	δ	
a) DEFICITS (PLURALITIES): differences in pluralities, from table 20(e)					
α	.56	.55	.59	.08	
β	.25	.44	.57	.10	
γ	.23	.38	.71	.12	
δ	.13	.22	.47	.50	
b) DIELEMENTARY ANALYSIS OF DEFICITS: Pluralities					
α	-.02	.14	.01	-.10	.07
β	.01	**-.17**	.06	-.01	.00
γ	.01	.06	-.05	.03	-.02
δ	-.06	-.07	.01	.07	-.05
	-.06	.04	.21	-.19	.30

and +.27 on the diagonal

c) DIELEMENTARY ANALYSIS OF DEFICITS: f-roots					
α	-.01	.12	-.01	**-.16**	.10
β	.03	**-.16**	.06	-.04	.00
γ	.01	.05	-.06	.02	.00
δ	-.09	-.05	.02	**.17**	-.03
	-.05	.04	.14	-.14	.37

and +.27 on the diagonal

d) DIELEMENTARY ANALYSIS OF DEFICITS: f-logs					
α	.02	.06	.00	**-.19**	.04
β	.03	**-.17**	.07	-.06	-.05
γ	.03	.04	-.08	.01	-.01
δ	-.07	-.05	.00	**.18**	.01
	.00	.02	-.03	.01	.60

and +.27 on the diagonal

table 23 (Cont'd)

e) SIZES OF STRIPE ADJUSTMENTS:

Stripe	Expression	α	β	γ	δ
Rows	**Plurality**	.07	.00	-.02	-.05
	f-root	.10	.00	.00	-.03
	f-log	.04	-.05	-.01	.01
Cols	**Plurality**	-.06	.04	.21	-.19
	f-root	-.05	.04	.14	-.14
	f-log	.00	.02	-.03	.01

f) SIZES OF RESIDUALS:

Expression	q-typical square	Typical midsquare
Plurality	.0039	.0053
f-root	.0106	.0048
f-log	.0072	.0053

In terms of f-log deficits, then, we can summarize the data quite well as follows:

typical deficit (diagonal) = +.87 = +.60 + .27

typical deficit (non-diagonal) = +.60

additional for $\delta\delta$ = +.18

additional for $\beta\beta$ = -.17

additional for $\alpha\delta$* = -.19

*father's cut at α, son's cut at δ.

Of the three situations with a distinctive residual, two are on the diagonal, suggesting that the diagonal effect may really be far from constant. We may, for instance, be well advised to give -- instead of one uniform diagonal effect and two large residuals -- all four individual diagonal deficits, perhaps as found from the dielementary analysis, namely:

for $\alpha\alpha$	.60 + .29 =	.89	(.93)
for $\beta\beta$	.60 + .10 =	.70	(.67)
for $\gamma\gamma$	.60 + .19 =	.79	(.75)
for $\delta\delta$	.60 + .45 =	1.05	(1.03)

(The values in parenthesis are the observed deficits.) To be sure whether this is a more useful description would require looking at much more extensive and varied mobility data.

The third far-out residual is of independent interest, relating as it does to professional (and semi-professional) fathers and unskilled sons. (These are the groups above α and below δ.) The natural hypothesis is that this lower barrier

reflects the "shirt sleeves to shirt sleeves in three generations effect" -- that it comes from unusually easy passage downwards from professional fathers to unskilled sons. How can we inquire as to whether this is so?

One natural way is to notice that each deficit is a difference -- see table 22 -- that the table of deficits is a difference of two tables, one for fathers occupation above the other for father's occupation below. The dielementary analysis of these two tables separately might help us to understand the situation. Table 24 gives these analyses. We see that the distinctive behavior of the NE corner cell is associated with father's occupation above α, and not with father's occupation below α. Accordingly, the deviation in the observed deficit is associated with an unexpectedly large fraction of unskilled sons of professional and semi-professional fathers. (Since only 15 sons were counted in this cell, we are not very sure how well this appearance would be likely to hold up.)

table 24 of chapter 13

Dielementary Analyses of the Two Tables of
Panel a of table 22

a) DIELEMENTARY ANALYSIS: Father's occupation above

Father's above	Son's occupation above α	β	γ	δ	
α	.11	.01	-.08	-.21	.25
β	-.01	-.03	.01	-.01	.04
γ	-.02	.00	.01	.09	-.03
δ	.01	.02	-.01	.00	-.35
	-.64	-.33	.33	1.80	-.09

and +.08 on the diagonal

b) DIELEMENTARY ANALYSIS: Father's occupation below

Father's below					
α	.06	-.03	-.01	.00	.13
β	-.07	.06	-.03	.01	.05
γ	-.09	-.09	.07	.05	-.05
δ	.03	.01	.00	-.26	.34
	-.02	-.32	.32	1.80	-.64

and -.15 on the diagonal

13Y. Problems

1. Display the data of table 6 in an effective picture. Comment on results. (What peculiarity must you be careful about?)

2. Reanalyze the data of table 7 symmetrically, using logarithms as in table 2. Comment on the difference between your results and those of table 7.

3. Analyze the data of table 8, using f-roots as suggested in the text. Comment on any regularities or irregularities you find.

4. Table 25 shows the after-career records of 223 delinquent boys divided according to their intelligence as measured by I.Q. tests. Analyze the table, comment.

5. Analyze one or more of the other tables given by David and Pearson at pages 64 to 66.

6. Carry through the calculations of table 17, keeping three decimal places in all logarithms.

7. The more complete version of the 1940 occupational mobility counts of Rogoff 1953 is as in table 26. Complete the calculations of tables 19 to 21 for this more detailed table.

8. Rogoff (1953) also gives counts for occupational mobility for sons observed in 1910. An extract is given as table 27. Analyze the table. Compare with the 1940 results.

table 25 of chapter 13

After-career Records of Delinquent Boys

Intelligence	Doing well	Fairly well	Badly	(T)
Normal and above	23	5	12	(40)
Dull normal	20	13	10	(43)
Border line	22	20	17	(59)
Mentally deficient	26	26	29	(81)
(T)	(91)	(64)	(68)	((223))

(T) = totals.

Source: F. N. David and E. S. Pearson 1961. Elementary Statistical Exercises. London; Cambridge University Press. (Esp. Table 63 on page 64.)

table 26 of chapter 13

Indianapolis 1940 Occupational Mobility Counts

a) CUTTING POINTS

Professional

ω

Semi-professional

α

Proprietors, managers, and officials

β

Clerical and sales

γ

Skilled

η

Semiskilled

δ

Unskilled

b) COUNTS

Son's occupation

	ω	α	β	γ	η	δ
ω	134	30	36	132	73	45	12
α	18	22	4	20	27	14	3
β	92	41	212	368	172	238	30
γ	84	57	83	461	165	179	26
η	89	78	117	520	880	734	154
δ	38	32	62	262	279	657	81
	17	11	20	94	111	216	206

table 27 of chapter 13

Occupation Mobility Counts for Indianapolis in 1910

a) CUTTING POINTS:

Same as Table 23 (a)

b) COUNTS:

Father's	Son's Occupation						
	ω	α	β	γ	η	δ	
ω	79	14	46	91	82	41	12
α	3	20	7	16	10	10	4
β	91	35	264	345	253	131	62
γ	37	19	49	288	145	70	24
η	51	43	111	414	1323	460	190
δ	23	14	34	129	299	297	97
	10	9	35	137	336	238	430

Source: Natalie Rogoff 1953. Recent trends in occupational mobility. Glencoe; Free Press. (Esp. Table 1 on page 44.)